Fonotaktická probabilita v češtině

Jan Henyš · jan.henys@ff.cuni.cz · Žďárek 2021

Fonotaktická probabilita

- Co to je?
- Jak se počítá?
- Jak vznikl nástroj pro počítání?
- Jak vypadá nástroj pro počítání?
- K čemu to všechno?

Fonotaktická probabilita

- **Fonotaktika**
  - Zabývá se **fonotaktickými komplexy** (fonotaktickým systémem), výskytem (distribucí), **kombinovatelností** a syntagmatickými vztahy fonologických jednotek uvnitř těchto komplexů
  - Fonotaktický komplex je uspořádaný svazek fonémů (fonotagma)
- **Probabilita** = pravděpodobnost

Fonotaktická probabilita

- **Fonotaktická probabilita odpovídá frekvenci výskytu fonologických segmentů (a sekvencí takových segmentů) v určitém jazyce.**

(Vitevich & Luce, 2004)

Kombinovatelnost uvnitř fonotagmatu

- Dichotomie legálních a ilegálních kombinací
  - Př.: /pl/ legální, /čř/ ilegální
  - `[word=".*čř.*"]` (syn 2020)

letní dvory i svátek už zívá prostor a jiné úhony	**počítačřadí**	ikony do oken světel mží v srpnu neohlíží se vyprávění
aby držely tvar . Spoustu vajíček , z dob ručních	**učřiu**	. prací mých dětí ještě na ZŠ , jsem připevnila
trochu „ tej domorodej kultúry a starích jejich kodexův a	**čřhův**	“ , bude po nás vědecky a morálně opodstatněná a
: Zkontrolujte si teplotu vždy , když budete mít pocit	**horečř**	ky . Může jít o první známku infekce nebo rejekce

- Fonotaktická probabilita jako **škála**

Počítání fonotaktické probability

Positional segment frequency

- Frekvence výskytu segmentu na pozici
- Př.: slovo plech
  - /p/ na 1. pozici
  - /l/ na 2. pozici
  - /e/ na 3. pozici
  - /x/ na 4. pozici

Biphone frequency

- Frekvence výskytu dvou po sobě jdoucích segmentů na pozici
- Př. slovo plech
  - /pl/ na 1. pozici
  - /le/ na 2. pozici
  - /ex/ na 3. pozici

Počítání fonotaktické probability

- Hall et al., 2015

	Original		
Word	**Trans.**	**Type Freq.**	**Token Freq.**
blick	[blɪk]	1	22
blep	[blɛp]	1	107
dwyk	[dwɪk]	1	3
mup	[mʌp]	1	57

Počítání fonotaktické probability

$$\text{PhonProb}_{\text{tokens}} = \left[\frac{\text{Sum of log frequencies of words with [bl] in initial biphone position}}{\text{Sum of log frequencies of words with any biphone in initial biphone position}} + \frac{\text{Sum of log frequencies of words with [lɪ] in second biphone position}}{\text{Sum of log frequencies of words with any biphone in second biphone position}} + \frac{\text{Sum of log frequencies of words with [ɪk] in third biphone position}}{\text{Sum of log frequencies of words with any biphone in third biphone position}} \right] / \left[\text{Number of biphone positions}\right]$$

...which in this specific case, translates to:

$$\text{PhonProb}_{\text{tokens}} = \left[\frac{log(22) + log(107)}{log(22) + log(107) + log(3) + log(57)} + \frac{log(22)}{log(22) + log(107) + log(3) + log(57)} + \frac{log(22) + log(3)}{log(22) + log(107) + log(3)} \right] / 3$$

$$= 0.43795$$

Počítání fonotaktické probability

- Hodnota FP se pohybuje **mezi 0 a 1**
  - U českých slov cca 0,01 — 0,001
- Praha × Žďárek

Nástroj pro počítání fonotaktické probability

- Input – **frekvenční seznamy** z korpusů ČNK
  - syn2020
  - oral v1
  - možnost přidání vlastního frekvenčního seznamu
- Zpracování frekvenčních seznamů
  - Čištění od nečeských slov (zkratky, nečeské znaky)
  - Převod do lowercase

Nástroj pro počítání fonotaktické probability

- Přepis do IPA pomocí knihovny corpy
  - Kontrola správnosti: 90 % pro syn 2020, 97 % pro oral v1
- Počítání OP/FP
- Spuštění přes command line v Pythonu

```
INFO:__main__:Generated ngram ('ʒ', 'ɟ') for position 0
INFO:__main__:Found 39 matches for ngram ('ʒ', 'ɟ') (334571),  matching_frequency_sum=58.852096324224675, all_frequency_sum=544867.2358609944
INFO:__main__:Generated ngram ('ɟ', 'a:') for position 1
INFO:__main__:Found 25 matches for ngram ('ɟ', 'a:') (333112),  matching_frequency_sum=39.31207746620122, all_frequency_sum=541968.0020149325
INFO:__main__:Generated ngram ('a:', 'r') for position 2
INFO:__main__:Found 159 matches for ngram ('a:', 'r') (327380),  matching_frequency_sum=246.75677559891682, all_frequency_sum=531827.8756858825
INFO:__main__:Generated ngram ('r', 'ɛ') for position 3
INFO:__main__:Found 1045 matches for ngram ('r', 'ɛ') (314183),  matching_frequency_sum=1634.9428315963078, all_frequency_sum=507645.4903221038
INFO:__main__:Generated ngram ('ɛ', 'k') for position 4
INFO:__main__:Found 1241 matches for ngram ('ɛ', 'k') (286317),  matching_frequency_sum=2071.312473558576, all_frequency_sum=457511.0083535731
INFO:__main__:Generated 5 ngrams
žďárek: 0.0016785030399530526
```

Souvislost s jinými metrikami

- **Neighborhood density**
  - Levenshteinova vzdálenost
    - Přidání znaku
    - Odebrání znaku
    - Změna znaku
    - Př. Slovo *kráva* (syn 2020, frekvenční práh f=5)
      - krávy, krávu, krávě, krála, krávo, krása, práva, káva, tráva
      - **ND = 9**
  - Předpoklad **pozitivní korelace FP s ND**

Souvislost s jinými metrikami

- Nižší reakční časy v **lexical decision task** u slov s **vyšší FP** (Luce & Large, 2001)
- Prezentace stimulu, respondent určuje, zda jde o *existující* slovo v určitém jazyce

K čemu to je?

- První podobný nástroj pro češtinu (slovanské jazyky?)
- Pseudoslova s **nízkou fonotaktickou probabilitou** jsou **zpracovávána pomaleji** (Janse & Newman, 2013)
- Hodnoty určování **wordlikeness** korelují s FP
- Možnost kontroly pro různé experimenty užívající slova jako stimuly

Reference

- Hall, K. C., Allen, B., Fry, M., Mackie, S., & McAuliffe, M. (2015). Phonological CorpusTools, Version 1.2. [Computer program].
- Janse, E., & Newman, R. S. (2013). Identifying nonwords: Effects of lexical neighborhoods, phonotactic probability, and listener characteristics. *Language and Speech*, *56*(4), 421-441.
- Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, *16*(5-6), 565-581.
- Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. Cognition, 78(2), 91-121.
- Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481-487.
- Vitevitch M. S., Luce P. A., Charles-Luce J., Kemmerer D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language & Speech,* 40, 47–62.